图书情报工作
第 64 卷 第 7 期 2020 年 4 月

春秋时期社会发展的主题挖掘与演变分析*
——以《左传》为例

■ 何琳 乔粤 刘雪琪
南京农业大学信息管理系 南京 210095

摘 要：[目的/意义]在人文计算迅速发展的背景下，利用文本挖掘技术对《左传》进行聚类计算，为春秋时期社会发展状况的主题挖掘等定量分析提供参考，同时对典籍文本多维度重组和分析也具有一定的借鉴意义。[方法/过程]采用文本聚类方法对《左传》进行多维度的定量分析，打破《左传》线性的编年体记载顺序，先运用词匹配算法从《左传》特征词语料中得到各个诸侯国语料，再将 LDA 主题模型先后用于处理《左传》特征词语料和选取的诸侯国语料，最后结合时间信息进行主题强度计算。[结果/结论] 实验结果表明，根据主题－词分布可以挖掘出春秋时期社会和诸侯国各方面的发展内容，通过主题强度变化曲线可以总结出春秋时期社会和各诸侯国的各方面发展态势。通过 LDA 主题聚类方法最终展现出了春秋时期整个社会以及不同诸侯国在战争、政治及外交等的发展变迁。

关键词：《左传》 主题挖掘 LDA 主题模型 主题演化 春秋时期社会变迁
分类号：G255
DOI：10.13266/j.issn.0252－3116.2020.07.004

1 引言

《左传》是先秦时期的重要典籍，是我国第一部编年体史书。它以《春秋》的记事为纲，以时间先后为序，记叙了上起鲁隐公元年(公元前 722 年)，下迄鲁哀公二十七年(公元前 467 年)，共 255 年的历史，记录了春秋时期 100 多个诸侯国政治、经济、军事、外交和文化方面的重要事件和重要人物，是研究中国先秦历史和春秋时期社会发展的重要文化材料。尽管《左传》整体以编年记事，但是实际事件与人物的复杂程度和逻辑关联远远超过了线性的解读能力，普通读者即使认真地按照文本顺序阅读，也会遇到很大的理解困难和记忆障碍。在这种情况下，借助计算机对相关信息进行处理将显著降低阅读的难度，提高信息获取效率与利用程度[1]。

以 LDA 模型为代表的主题模型是一种广泛应用于文本分析的聚类方法，本文拟利用主题模型和自然语言处理技术对《左传》进行主题聚类，结合时间信息绘制各主题强度的变化曲线，分析春秋时期战争、政治、礼仪、外交等方面的发展与变迁。该方法打破《左传》线性的编年体记载顺序，通过不同的主题维度展现春秋时期整个社会以及不同诸侯国在战争、政治及外交等方面的发展变迁，以为春秋时期的主题挖掘等定量分析提供参考与借鉴。

2 相关研究综述

2.1 文本计算的相关研究

随着大数据时代的到来，结合计算机技术对大规模历史文化资料进行定量分析已成为人文研究中一种新的有效方法，通过统计分析从大规模数据中挖掘新事实、产生新认识[2]，能够发现靠传统文献阅读无法发现的隐藏在文献中的重要内容。这类研究多采用数据可视化、词频统计、自动分类聚类和机器学习等计算机技术对文本进行量化分析手段。如 TextArc 文本可视化分析工具针对单一文本，将目录、概要、叙词表和词频统计共现结合在一起，从整篇文档和文档中的词两

* 本文系国家社会科学基金项目“基于典籍的中华传统文化知识表达体系自动构建方法”(项目编号：18BTQ063)研究成果之一。
作者简介：何琳(ORCID：0000－0002－4207－3588)，教授，博士，博士生导师，E-mail：helin@njau.edu.cn；乔粤(ORCID：0000－0002－1968－9608)，硕士研究生；刘雪琪(ORCID：0000－0002－5346－7291)，硕士研究生。
收稿日期：2019－07－10 修回日期：2019－10－26 本文起止页码：30－38 本文责任编辑：易飞

何琳，乔粤，刘雪琪. 春秋时期社会发展的主题挖掘与演变分析——以《左传》为例[J]. 图书情报工作,2020,64(7):30-38.

个方面对全文本进行交互式可视化计算分析,引导读者揭示新发现[3]。T. Horton 等针对所选择的部分 19 世纪美国小说,先按照高、中、低的情感强度标记小说中的每一章节,然后利用机器学习测试新的小说文本,挖掘出代表不同情感强度的具体词汇和情境[4]。F. Moretti 在 2010 年建立斯坦福文学实验室,通过自动聚类和机器学习,从多风格的文学作品中归纳出人无法总结的文本模式特征,然后对未知样本进行分类实验,后续研究结合网络分析理论。他们把《哈姆雷特》的情节用网络关系表达出来,分析人物对白等关系等,以帮助探究文艺复兴时期欧洲国家的发展[5]。J. B. Michel 等选择了 500 多万本 1800 年至 2000 年出版的英文书籍,通过计算词语的共现频率和统计分析,归纳出"技术被采纳的速度越来越快"等相关研究结论[6]。

2.2 典籍文本的挖掘研究

近年来随着文本计算相关技术的发展,针对中国典籍进行文本实验的研究也逐渐增多。J. W. Chen 等学者对《世说新语》进行文本内容分析,使用柱状图、Gephi 网络关系图、GIS 地理信息图等方法对原书主题、人物以及包含的地理空间信息进行了挖掘与量化分析[7]。美国加州大学洛杉矶分校的"东亚研究巨视显微镜"(East Asian Studies Macroscope,EASM)项目,针对《全唐诗》利用主题建模聚类等方法进行处理,提炼出一套文档内的潜在语义模式,用户可以通过交互界面分析诗歌主题、内容等[8]。欧阳剑对古籍文本进行字词的历时词频分布规律可视化分析,以中国史定量研究为例,对部分中史的经典宏观理论从量化角度进行了初步验证,例如分析了古籍文本中对武则天的评价和"重学轻术"思想的影响等[9-10]。印第安纳大学和西安交通大学合作开发的工具 InPhO Topic Explorer,将概率主题建模应用到其建立的 Handian(汉典古籍)语料库上,用来辅助发现和解释该语料库中的主题[11]。R. Nichols 等对中国古代和中世纪的 500 多万字的语料库进行主题建模,根据文本的相交主题和不相交主题,解释验证了中国古代哲学的重要文本《论语》《孟子》和《荀子》之间的竞相关系[12]。

2.3 《左传》文本的相关研究

目前针对《左传》的研究,大部分都是具有人文学科背景的研究者利用详细阅读等一些人文学科的研究方法进行研究,主要集中于字句注疏的考证、《左传》与先秦文献关系及成书时间考证、《左传》内容及思想等[13]。《左传》内容及思想方面的研究相对比较分散,主要有战争[14-15]、礼仪[16-17]、人物[18-19]、外交[20-21]等方面,呈现出单一线性、不立体的特点,其中《左传》词汇与句子相关研究也均是大量人工整理所得,这种方法不利于大规模古籍文本研究。而采用文本计算的方法对《左传》进行挖掘的研究还比较少,其中许超、陈小荷提取了《左传》中的人物与事件,使用社会网络分析软件 Pajek,建立起春秋时期社会网络,定量地对这一时期的历史社会网络做了探索性研究,得出了依靠传统研究方式难以实现的一些结论[22]。这类定量化的研究较少的原因在于在运用现代计算机技术方法处理古籍文本时,典籍文本自身存在的一些独特属性往往加大了研究难度。以《左传》为例,一是其原始的段落划分太过细致,大部分段落文本篇幅过短且语义不集中;二是中文词语是高度多义的,不同的词性所代表的含义也是多样的;三是《左传》的编年体记述方式,只是按照年份顺序记述发生的事件过程,同一年份中各方面的事件相互掺杂,主题分散。因此,加大了文本处理与分析技术的难度。

综上可知,机器学习等文本分析技术的发展为利用计算机技术对大规模历史文化资料进行定量分析提供了坚实的技术基础,学者们也成功地将文本分析技术用于《全唐诗》《汉典》等典籍的挖掘。《左传》是一本研究春秋时期的重要典籍,长期以来学者们进行了大量定性的研究和分析。然而《左传》以编年体方式记载,庞杂的主题、人物、事件交杂在一起,给人工分析带来了巨大的挑战。在人文计算研究发展如火如荼的今天,利用文本分析技术对《左传》进行细粒度的主题挖掘研究,对于深度挖掘典籍中隐含的知识具有重要的意义。本文正是在前人研究的基础上,针对《左传》文本自身的属性特征,如文本聚类粒度、古文特征词提取方法和主题挖掘的研究维度等方面存在的问题,对现有的主题挖掘技术进行了有针对性的优化。使用经过优化后的主题建模与主题演化的方法对《左传》文本进行了多维度的聚类计算分析,研究结果对于典籍文本多维度重组和分析具有一定的参考借鉴意义。

3 研究方法

3.1 研究框架

据史学知识所知,春秋时期社会的演变发展主要围绕在战争、外交、政治和礼仪等几个方面,本文分别从整个社会及各个诸侯国两个维度出发,首先通过各自的主题建模得到主题聚类的结果,然后结合时间信息对主题建模的结果数据做进一步处理,得到主题强度变化图,通过主题演化研究来挖掘春秋社会 255 年

间各时期不同主题的演变态势及发展变迁。

本文的研究框架如图 1 所示，首先选取古籍《左传》的初始语料，对其进行数据预处理，然后进行主题建模预实验以确定聚类粒度即文本单位的长短，加以整理得到《左传》特征词语料。运用词匹配算法，从《左传》特征词语料中得到各个诸侯国语料，选取代表性诸侯国的语料进行下一步处理。将 LDA 主题模型先后用于处理《左传》特征词语料和选取的诸侯国语料，根据所得《左传》主题－词分布来挖掘出春秋社会发展的主要方面和内容，在此基础上，根据各诸侯国的主题－词分布来挖掘它们各方面的发展内容。最后结合时间信息，对得到的《左传》和各诸侯国的文档－主题分布进行主题强度计算，通过得到的主题强度变化曲线反映出各方面的发展演变状况，总结春秋时期社会和各诸侯国的各方面的发展态势，最终展现春秋时期社会各方面的发展变迁。

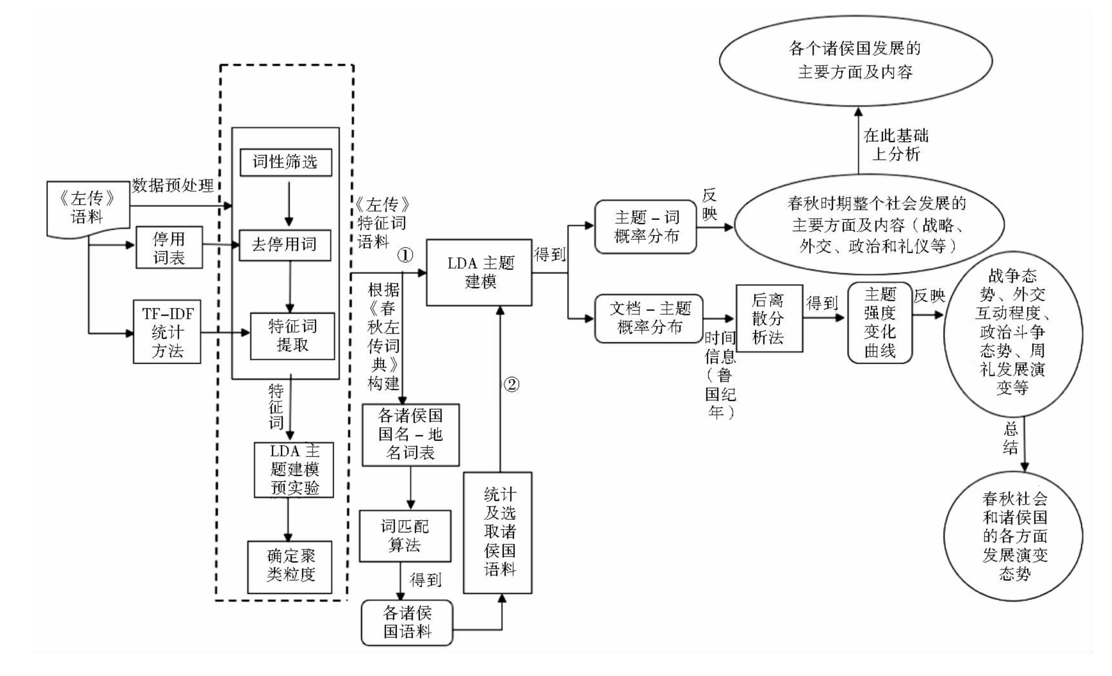


图 1 春秋时期社会发展主题挖掘及演变研究框架

3.2 关键步骤实现方法

3.2.1 LDA 主题建模

LDA(Latent Dirichlet Allocation)[23]是 D. M. Blei 于 2003 年提出的三层贝叶斯主题模型，它通过无监督的学习方法发现语料库中隐含的主题信息，采用词袋(bag of words)的方法将语料库中的每一篇文档视为一个词频向量，从而将文本信息转化为易于建模的数字信息。由于 LDA 模型的原理与算法实现较为普遍成熟，本文在此不做过多的介绍。本文将利用《左传》特征词语料的主题建模结果之一——主题－词概率分布来获取春秋社会发展的主要主题维度和具体内涵。在此研究基础上，将选择的诸侯国语料进行同样的主题建模，利用各诸侯国的主题－词分布来挖掘和定义各诸侯国的不同主题聚类维度的发展内容。

在 LDA 主题建模中，主题数量的选择对于结果分析非常重要。太少的主题可能会把语义上不相关的内容合并到所谓的嵌合主题中；太多可能会导致相关内容分裂成单独的主题，造成主题之间的冗余或不相关的“垃圾”主题的积累[24]。主题质量通常由主题中关键词的语义一致性决定，但主题一致性的评估通常是研究者解释的产物[25]。因此，本研究在使用 Python 调用第三方的 LDA 库时，对《左传》特征词语料进行了不同主题数目的若干试验尝试，经过分析最终确定主题数目为 6，迭代次数为 1 500，获得最优的主题识别结果，这些主题提供了“广度”和“深度”的良好平衡，同时产生了很少的垃圾或嵌合主题。

3.2.2 基于 LDA 的主题演化

主题会随着时间进行相应的变化，引入时间因素，这种变化通常反映在强度、内容两方面，主题强度描述了一个主题的受关注程度，通常也代表着主题的热门程度，本文通过观察主题强度随时间的不断变化来掌握主题演化的方向。在主题演化研究方面，T. L. Griffiths 等[26]在 2004 年首先提出了后离散分析方法，即先在整个文档集上用 LDA 主题模型获取所有的主题，进而估计出 LDA 模型的参数，然后将文档离散到相应的时间片，对于某个主题依次计算它在每个时间片的

强度，从而显示了随时间推移而强度明显上升的热话题(hot topic)和下降的冷话题(cold topic)。

具体步骤为：

(1)划分时间片。《左传》以鲁国纪年为顺序，本文即按照鲁国十二公的顺序，将主题建模所得的所有文档－主题分布均划分成12个时间片。

(2)主题强度计算。主题建模中增量吉布斯抽样算法依据相应公式获得每个文档中主题的概率分布，即θ分布(文档－主题分布)，本文根据这个分布利用崔凯[27]提到的主题强度衡量方法来计算主题强度，计算公式为：

$$T_K = \frac{\sum_i^M \theta_{ki}}{M} \qquad 公式(1)$$

把文档扩展到文档集，M为时间片中文档集的总数目，θ_{ki}表示主题k在第i篇文档中出现的概率，T_K为主题K在该时间片中的平均概率，即主题K的强度值，通过重复计算可以得到主题K在12个时间片中不同的强度值，每个主题经过上述相同的计算最后可绘制出所有主题的强度变化曲线图。

(3)定量分析。根据主题强度变化曲线图，结合此前相应的主题－词分布的语义分析，量化分析每个主题的演化情况，分析并总结春秋时期整个社会和各个诸侯国的各方面的发展态势。

3.2.3 特征词抽取

特征词的抽取影响LDA主题聚类的效果。根据此前本节描述的《左传》文本的特点，为了提升主题聚类的效果，本文的特征词抽取方案如下：

(1)构建基础关键词典。本文的训练语料是经过手工分词得到的《左传》语料，该语料是南京师范大学陈小荷研究团队的成果，由具有古文知识背景的研究生手工切分完成。由于已知动词语义表达的重要性，因此与处理现代汉语语料不同的是，除了保留蕴含语义最强的名词与形容词，本文还提取了动词作为基础关键词。

(2)去除停用词。为了提高主题学习的质量，需要过滤一些不能反映文本主题的词语，本文的停用词表采用的是汉典古籍停用词表[11]，它是针对中国古籍文本而由人工制作的，包含187个词语。

(3)基于TF-IDF的特征词抽取方法。至此，语料的无效干扰词还有很多，而特征词的选取可以进一步删除其中的干扰词以提高LDA模型输入数据的有效性[28]。TF-IDF是一种统计方法，用以评估一个词语对于一个语料库中的其中一份文档的重要程度，TF-IDF实际上是TF(词频)＊IDF(逆向文件频率)，其具体计算公式为：

$$w_{i,j} = tf_{i,j} \times \log\left(\frac{N}{df_i}\right) \qquad 公式(2)$$

其中，$w_{i,j}$表示词语i在文档j中的TF-IDF值，$tf_{i,j}$表示词语i出现在文档j中的频率，N表示语料库中文档总数，df_i表示在整个语料库中包含词语i的文档数。对经过(1)(2)步处理的《左传》语料进行TF-IDF的计算，选取TF-IDF值在某一区间中的词汇作为特征词，最终决定删除TF-IDF值低于特定阈值2的文本代表意义不大的无效干扰词，至此得到《左传》特征词语料。

3.2.4 文本聚类粒度

文本聚类粒度即文档长短也是影响LDA模型主题建模效果的重要因素，短文本代表性词少、文档级的词共现稀疏的特点使得LDA对于短文本的主题挖掘不一定能够达到理想效果。

因此，我们提出了3种聚类粒度的方案：

(1)以《左传》原始段落为聚类单位。根据此前分析得知《左传》文本的原始段落大都较短，可能并不适合直接用于LDA主题建模。

(2)以《左传》年份为聚类单位。例如隐公元年这一年份的所有文本为一个聚类单元，但这一年份中，存在军事、政治、外交等多个主题，可能导致聚类结果不佳。

(3)以《左传》年份中每个主题段落为聚类单位。主题段落是指文本中语义相近的多个原始段落作为一个聚类单元。本文采用中华书局《左传》(郭丹译文)中历史学专家划分的主题段落，将其作为文本聚类单位。

经过对3种聚类方案的对比，发现相对于全文文本和原始段落文本，方案(3)主题段落文本的聚类结果语义更集中明确，最适用于LDA主题建模。

3.2.5 诸侯国语料的获取

之前提到《左传》文本的编年体记述所带来的弊端，因此研究春秋时期社会发展还需要分析诸侯国层面的发展。为了获取不同诸侯国的主题语料，首先根据《春秋左传词典》(杨伯峻版)[29]整理的《左传》中158个诸侯国的国名－地名词表，对《左传》特征词语料中的每个主题段落进行国名和地名遍历，如果含有诸侯国词表中的任意一个国名或地名，则判定该主题段落属于此国名或地名所属的诸侯国语料。通过对诸侯国语料的统计分析，可得到每个诸侯国所涉及的主

题段落数，作为后续以诸侯国为主题分析维度的依据。表 1 显示了按照所含主题段落数降序排列的前 30 个诸侯国主题段落的数目。

表 1　不同诸侯国文本主题段落分布数量

（单位：个）

国家	主题段落数	国家	主题段落数	国家	主题段落数
魯国	811	燕国	175	蔡国	88
晉国	717	許国	175	申国	82
鄭国	621	曹国	158	向国	81
齊国	577	秦国	141	貳国	71
楚国	528	道国	128	韓国	66
宋国	453	過国	122	夏国	65
衞国	414	莒国	105	越国	64
陳国	290	夷国	104	梁国	63
吴国	198	陽国	101	州国	61
邾国	193	共国	96	隨国	61

4　实验结果与分析

本部分将按照第三部分的研究框架，运用其中的实现方法分别对其涉及的方面进行处理分析，并对实验结果进行对应的春秋时期社会和诸侯国的主题挖掘与演变分析。

4.1　春秋时期社会整体分析

如表 2 所示，《左传》主题－词概率分布描述了每个主题的概率排名前 15 个关键词，此前的特征词提取方法，使得这些结果词大都具有高度语义，代表了 LDA 主题模型关于《左传》中整个春秋时期社会的 6 个重要主题的发现。具体分析如下：

表 2　《左传》主题－词概率分布

主题 0	概率	主题 1	概率	主题 2	概率	主题 3	概率	主题 4	概率	主题 5	概率
晉	0.035 456	王	0.029 251	齊	0.029 013	楚	0.041 842	死	0.023 663	無	0.026 576
諸侯	0.027 242	書	0.012 634	齊侯	0.023 936	師	0.040 558	立	0.021 972	可	0.022 165
盟	0.021 626	名	0.012 179	魯	0.019 463	鄭	0.020 922	殺	0.021 458	民	0.021 607
宋	0.018 776	周	0.011 951	衞	0.016 441	陳	0.018 505	臣	0.018 519	能	0.018 034
晉侯	0.016 094	日	0.011 155	取	0.013 661	王	0.017 296	告	0.015 506	德	0.016 080
命	0.015 759	生	0.009 675	齊人	0.010 155	晉	0.016 768	罪	0.014 624	國	0.015 410
寡君	0.013 328	火	0.009 561	莒	0.009 672	敗	0.016 390	命	0.012 640	禮	0.010 887
子產	0.012 993	疾	0.008 992	季孫	0.009 672	伐	0.015 861	奔	0.011 023	對	0.010 832
辭	0.012 490	物	0.008 196	城	0.009 309	吳	0.013 218	生	0.009 407	亡	0.009 380
鄭伯	0.011 316	晉侯	0.008 082	伐	0.009 188	戰	0.012 614	公子	0.007 717	政	0.008 933
會	0.011 149	祀	0.008 082	門	0.009 188	楚子	0.012 614	大夫	0.006 982	失	0.008 263
禮	0.010 814	食	0.007 854	成	0.009 067	秦	0.012 538	車	0.006 688	詩	0.008 152
朝	0.010 646	官	0.007 513	邾	0.008 100	救	0.010 272	召	0.006 467	行	0.008 040
大夫	0.009 975	時	0.007 399	叛	0.007 254	帥	0.010 197	歸	0.006 320	善	0.007 928
衞	0.009 808	戎	0.006 944	圍	0.007 133	許	0.010 046	舍	0.006 320	天	0.007 370

主题 0 中大量动词“盟”“命”“辭”“會”“朝”“衞”均为朝聘会盟类动词，其中会盟类代表词“盟”概率排名第三。周王室虽衰败，但周礼在当时还是个庞大的精神存在，因此会盟其实是表面上尊奉周天子，本质是调遣不臣服的诸侯，这一点由排名第 2 的“諸侯”、第 12“禮”等关键词体现，其他名词关键词均为天子、诸侯和臣子。因此主题 0 的标签定义为“诸侯会盟”。

主题 1 中概率排名靠前的“王”“周”代表周天子周王朝，“書”（记载），“名”（爵号），“官”“日”（日月星辰），“生”（出生），“火”（点火仪式），“疾”（疾病），“物”“祀”“食”（祭祀時配享、配食等），代表封官加爵，生辰与疾病，记录日月星辰之轨迹，祭祀的食品、物件及仪式，礼仪贯穿于春秋社会的方方面面。因此主题 1 的标签定义为“礼仪、仪式”。

主题 2 中的“取”（战胜而获取；以强力夺取他人之物；灭人之国以扩张己地），“伐”（征伐），“成”（成功，讲和，调解使和），“叛”（背叛），“圍”（包围）大量的关键动词揭示了春秋时期是强者为尊，以实力争霸的时代以及各诸侯国之间错综复杂的关系。其中实力较强的诸侯国相关词概率排名相对靠前，有“齊”“齊侯”“魯”“衞”，鲁臣“季孫”，小诸侯国有“莒”“邾”，一定程度揭示了这些诸侯国及其诸侯臣子之间的复杂关系。因此主题 2 的标签定义为“诸侯国关系”。

主题 3 中的“師”“敗”“伐”“戰”“救”“帥”均为战争类动词，表明战争是春秋时代的显著元素。该主题还体现出很多诸侯国“楚”“鄭”“陳”“晉”“吳”“秦”

“许”之间讨伐与营救的错杂关系。因此主题3的标签定义为“诸侯国战争”。

主题4中动词“死”“立”(立人为君;生存),“殺”“告”(告诫;宣告),“罪”“命”(命令),“奔”(奔逃,出走,流亡),“生”(出生,使复生),“召”“歸”“舍”(任命,安置)以及人物名词“臣”“公子”“大夫”揭示了春秋时期政治发展的新格局:争霸的重心从诸侯国之间转移到各诸侯国内位高权重的卿相大夫之间,各国内乱迭起。因此主题4的标签定义为“诸侯国的宫廷权力斗争”。

主题5中的“民”“國”“政”表明在国家的政治治理中人民的重要性,其中“民”,概率排名第三,体现出人本思想;排名靠前的“可”(表示赞同),“能”(才能),“德”“禮”表示《左传》中倡导赞同国君要有才能,有德行,以礼治国;“詩”(诗经),“行”(奉行,执行),“善”“天”,则表示《左传》中倡导国君要有善心,信天命,同时也体现出《左传》对诗经的引用;“無”(表示否定),“亡”(被消灭),“失”(违背,丢失)这类词传达的是如果国君背弃了这些仁义道德,那么就会被灭亡的思想。因此主题5的标签定义为“周礼治国”。

本文的文本聚类粒度为主题段落文本,主题段落语义明确集中,在此基础上得到的文档－主题概率分布更为明确,计算得出的每个时间片的主题强度能够准确地体现出主题的发展热度。以时间(鲁国十二公)为横轴、主题强度为纵轴绘制部分主题强度变化图,如图2所示:

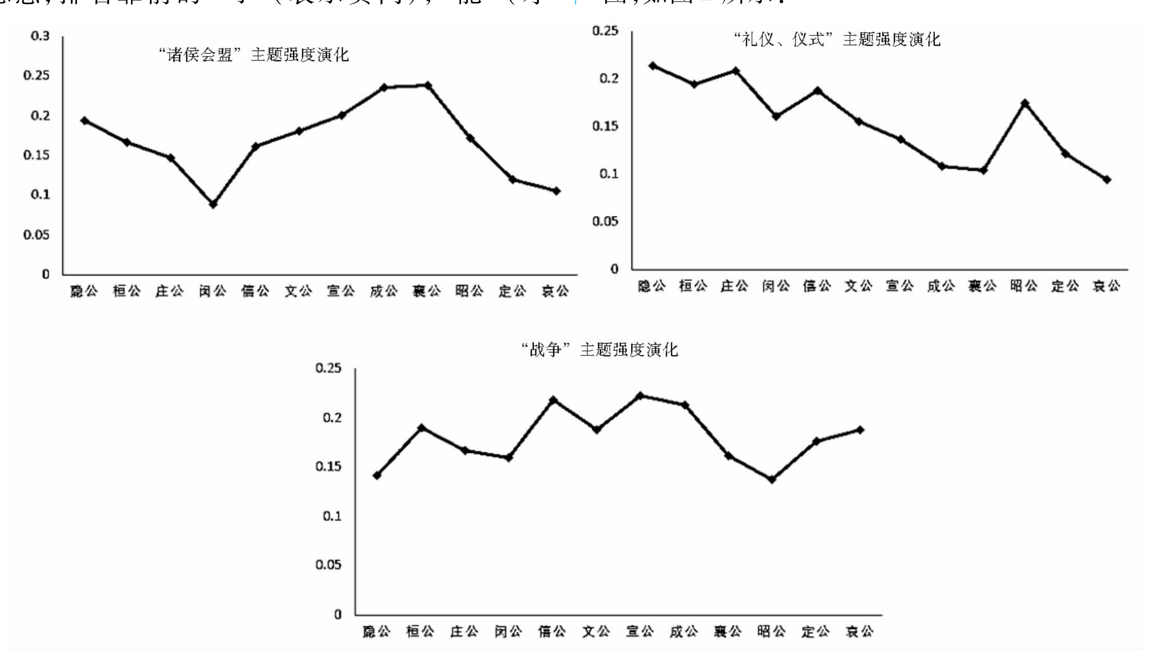


图2 《左传》春秋社会部分主题强度变化

主题0“诸侯会盟”在闵公时期热度最低,此后持续高涨,直到襄公时期开始下降。春秋初期,周王室逐渐没落,会盟诸侯也逐渐减少,直到齐桓公以“尊王攘夷”为号召,会盟诸侯,逐渐变成霸主,即正是闵公时期诸侯会盟开始逐渐回升增加。春秋中晚期政治格局出现的新趋势:争霸的重心从诸侯国之间转移到各诸侯国内位高权重的卿相大夫之间,各国内乱迭起,因此诸侯国之间的会盟的热度开始从襄公时期下降。

主题1“礼仪、仪式”一直贯穿于整个春秋时期,但是随着周王室的没落,礼乐制度土崩瓦解,热度一直逐渐下降。虽然鲁襄公到鲁昭公时期有小回升,原因可能是因为鲁襄公时期孔子出生,他推崇礼仪起了一定作用。

主题3“诸侯国战争”主题热度一直处于中上游水平,证明了战争在左传中的重要性。此前提到齐桓公“尊王襄夷”,从那开始会盟与征伐的交替进行就成为春秋时期的显著特征,因此闵公时期之后一段时期战争主题热度较高。公元前546年(鲁襄公时期)第二次“弭兵之会”达成协议,战火暂时得以平息,可以看出襄公时期之后战争热度开始下降。

4.2 春秋时期重点诸侯国分析

由第三部分的统计分析可知鲁国、郑国、楚国3个诸侯国在《左传》语料库中涉及的篇幅较大,所属的主题段落数均超过500,且在人文学科对春秋社会的研究中也很重要,因此在诸侯国维度上选择这3个诸侯国作为研究代表,延续《左传》特征词语料的主题建模

处理步骤,将 3 个诸侯国的语料分别进行主题建模处理,即使用 Python 调用第三方的 LDA 库,主题数目设置为 6,迭代次数设置为 1 500,得到表 3 – 表 5 三个诸侯国的主题 – 词分布,每个主题的前 15 个关键词(按照主题内的词语概率降序排列),在上一节对《左传》中春秋社会的主题 – 词概率分布语义分析的基础上,解读每个诸侯国的主题 – 词分布,并将每个诸侯国的主题加以标签。

表 3 鲁国主题 – 词分布

主题序号	主题关键词	标签
0	齊 齊侯 魯 門 奔 臣 季孫 死 取 告 城 莒 書 齊人 立	与齐国的关系
1	師 楚 戰 王 敗 陳 伐 吳 楚子 帥 軍 鄭 可 侵 獲	战争
2	民 無 能 可 國 德 禮 對 失 行 大夫 命 詩 善 心	周礼治国
3	王 秦 晉 晉侯 取 戎 周 立 伐 秦伯 大子 田 聽 狄 德	与其他诸侯国的关系
4	晉 諸侯 宋 鄭 盟 許 子產 會 寡君 辭 晉侯 齊 大夫 成 禮	诸侯会盟
5	死 殺 臣 罪 生 命 告 立 歸 卒 可 食 過 無 召	宫廷权力斗争

表 4 郑国主题 – 词分布

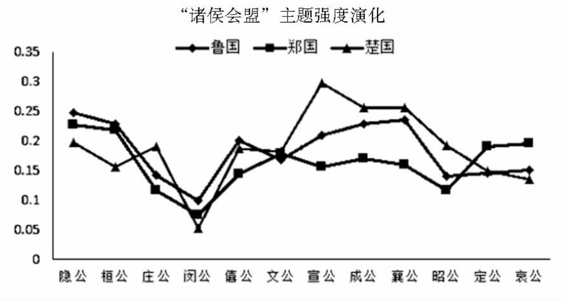
主题序号	主题关键词	标签
0	死 殺 臣 立 告 罪 奔 亡 止 取 歸 亂 舍 書 大夫	宫廷权力斗争
1	無 國 能 晉 可 子產 命 行 對 寡君 民 求 禮 難 朝	政治治理
2	生 晉侯 命 天 秦 火 戎 立 狄 祀 吉 民 大子 卜 名	礼仪、仪式
3	王 德 民 周 禮 亂 親 可 詩 取 行 威 舉 克 爭	周礼治国
4	楚 師 鄭 伐 晉 敗 陳 可 戰 楚子 帥 軍 許 侵 救	战争
5	盟 諸侯 齊 宋 晉 魯 會 命 齊侯 許 告 歸 晉侯 鄭伯 書	诸侯会盟

表 5 楚国主题 – 词分布

主题序号	主题关键词	标签
0	殺 死 立 臣 罪 生 無 歸 奔 公子 納 亡 告 書 王	宫廷权力斗争
1	晉 楚 鄭 諸侯 盟 許 陳 國 宋 無 蔡 歸 晉侯 鄭伯 楚子	诸侯会盟
2	师 楚 王 敗 吳 伐 戰 軍 帥 楚子 克 令尹 奔 秦 備	战争
3	王 無 德 民 周 戎 親 心 滅 成 神 疾 聽 封 啟	礼仪、仪式
4	齊 死 魯 齊侯 門 告 衛 乘 止 免 車 難 季孫 執 馬	与其他诸侯国的关系
5	可 命 能 民 行 對 臣 棄 禮 辭 天 求 失 國 政	周礼治国

其中,鲁国主题 0 中“齊”“齊侯”“齊人”体现出齐国在鲁国发展中的重要地位以及两国之间复杂的关系,成公年间,鲁国的卿家贵族“季孫”听盟主晋国号令,发兵攻齐,是为鞍之战。郑国主题 1 概率排名靠前的关键词“子產”是郑穆公之孙,在执政期间,进行了自上而下的改革,在对楚、晋等国的外交方面,也取得了一定的成绩,再结合其他关键词分析得出该主题的标签“政治治理”。楚国主题 2 表明其战争主要发生在“吳”“秦”两国之间,春秋时期吴国、楚国、秦国均为中原争霸的重要参与国家,吴王阖闾时期还攻破了楚国都城,“楚”“吳”排名靠前则体现出吴国和楚国之间较强的战争关系。

3 个诸侯国均有“诸侯会盟”和“战争”等主题,图 3 为绘制的相同主题的强度变化图。



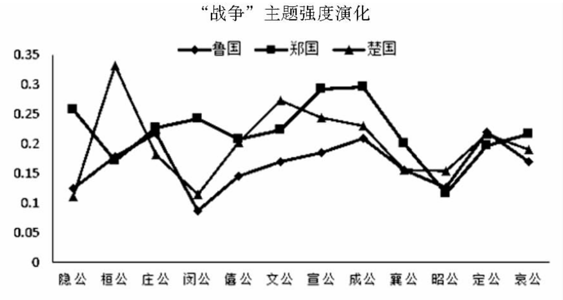


图 3 三国共同主题强度变化

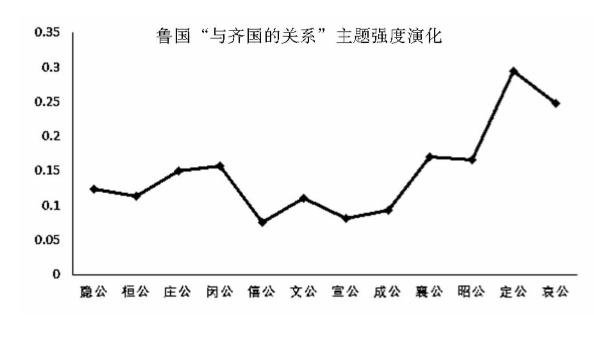
3 个诸侯国的“诸侯会盟”主题强度走势几乎吻合,整体来看主题热度处于中上游水平,说明诸侯会盟发生得较频繁,尤其是在闵公时期之后上涨幅度较大,这印证了上一节分析的整个春秋社会的“诸侯会盟”的演变态势。

春秋时期大大小小的战争不断,各个国家的状况都不同,其中由于鲁闵公仅在位两年,该期间鲁国几乎未参与战争,因此闵公时期鲁国“战争”主题热度最低。郑国“战争”主题强度起伏较大,在宣公、成公时期,郑国依违于晋、楚两大国之间,先后发生过很多次战争,并攻打了秦国、许国等,因此这两个时期的“战争”主题热度达到最高。鲁桓公时期是楚国势力扩张的时期,先后讨伐随国,与巴国一起夹攻邓国,进攻绞国等,战争讨伐不断,因此桓公时期楚国“战争”主题热度最高。

某些诸侯国还存在独自特有的主题,如鲁国“与齐

国的关系”主题、郑国“政治治理”主题，见图 4。

鲁国“与齐国的关系”主题热度在整个春秋时期逐步上升，在定公时期主题热度达到最高点。鲁定公在位期间被孔子建议外联齐国，制定了一系列措施，定



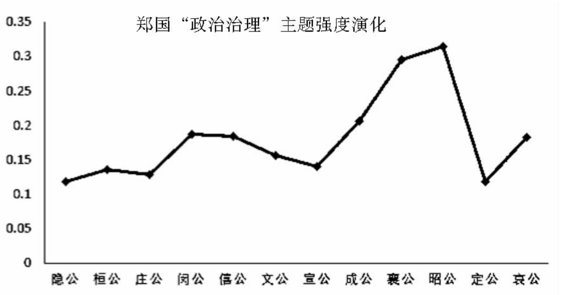


图 4 单个国家特有主题强度变化

公十年，在齐鲁夹谷之会鲁国又从齐国手中讨回了汶阳之地，齐鲁关系发展到最高热度。郑国“政治治理”主题热度变化起伏较大，在襄公和昭公时期热度最高，结合此前对郑国“政治治理”主题的分析，郑穆公之孙子產在公元前 554 年为卿，公元前 543 年执政，其执政期间，进行了自上而下的政治改革，对应鲁国纪年正是襄公和昭公时期。

5 结论与展望

主题模型为人文科学中的计算机辅助解释提供了一个有效的方法。本研究在近年国内外研究的基础上，以《左传》为例，针对古文进行特征词提取方法的初步探究，构建诸侯国国名 – 地名词表，结合文本挖掘技术进行国家语料的抽取与统计，运用 LDA 主题模型和主题演化方法，从《左传》中整个社会和各个诸侯国两个维度来挖掘春秋时期社会发展各方面的内容与演变态势。针对整个春秋社会分析，发现其主要围绕“诸侯会盟”“礼仪、迷信”“诸侯国关系”“诸侯国战争”“诸侯国的宫廷权力斗争”“周礼治国”这六大主题发展，在此基础上，分析发现各大诸侯国的发展也均围绕着“诸侯会盟”“战争”“宫廷权力斗争”等主题，其中还发现了某些诸侯国特有的发展主题，例如郑国的“政治治理”等，通过绘制的主题强度变化图清晰地描述出春秋时期整个社会和各大诸侯国的各方面发展随鲁国时间的热度变化，探索春秋时期社会各方面发展的变迁，证实了人文计算领域可借助 LDA 主题模型来理解、探索和诠释中国丰富的文化遗产，具有一定的实践意义。

本文的主题挖掘及演化研究方法有许多局限性。首先，虽然机器学习的方法不存在人为偏见，但是古典汉语中广泛的多义性给非历史学专业出身的笔者提供了解释性挑战，对主题内容与演化的分析解释可能会存在一定偏见与疏漏。其次，本研究发现古代汉语动词语义网络体系的构建与古籍主题段落的有效划分是影响主题建模效果的两个重要因素，由于时间精力有限，本研究还未对其进行更为深入的探究，在后续研究中将进一步进行探索。再者，人文计算的主题模型没有用于评估的正确的“黄金标准”[11]，后续研究中可以建立一个有效评估模型实验效果的评价体系，例如可以采用定量(困惑度等)与实验文本分析(专业人士的验证检查)相互结合的评价方法。最后，主题演化研究包括主题内容和主题强度两个方面，主题强度的演化衡量的是主题受关注程度的变化，主题内容的演化衡量的是主题关注点的迁移，本文的主题演化研究考量的是主题发展的热度的变化，而主题内容随时间的具体变化对于深入探索春秋时期社会各方面的发展变迁也非常重要，这亦是笔者下一步的研究工作。

参考文献：

[1] 胡悦融，马青，刘佳派，等. 数字人文背景下“远距离可视化阅读”探析[J]. 图书馆论坛,2017,37(2):1 – 9.
[2] 梁晨，董浩，李中清. 量化数据库与历史研究[J]. 历史研究，2015(2):113 – 128,191 – 192.
[3] PALEY W B. TextArc:showing word frequency and distribution in text [C]// IEEE symposium on information visualization. Poster Compendium: IEEE CS Press,2002:148 – 165.
[4] HORTON T, TAYLOR K, YU B, et al. “Quite right, dear and interesting”: seeking the sentimental in nineteenth century American fiction [EB/OL]. [2019 – 06 – 20]. http://www.csdl.tamu.edu/~furuta/courses/06c_689dh/dh06readings/DH06-081-082.pdf.
[5] MORETTI F. Distant reading[M]. London: Verso Books,2013: 211 – 221.
[6] MICHEL J B, SHEN Y K, AIDEN A P, et al. Quantitative analysis of culture using millions of digitized books[J]. Science, 2011, 331(6014): 176 – 182.
[7] CHEN J W, BOROVSKY Z, KAWANO Y, et al. The Shi Shuo Xin Yu as data visualization[J]. Early medieval China, 2014(S0):23

-59.
[8] CHEN J W. East Asian studies macroscope [EB/OL]. [2019-06-20]. http://macroscope.cdh.ucla.edu.
[9] 欧阳剑. 大规模古籍文本在中国史定量研究中的应用探索[J]. 大学图书馆学报,2016,34(3):5-15.
[10] 欧阳剑. 面向数字人文研究的大规模古籍文本可视化分析与挖掘[J]. 中国图书馆学报,2016,42(2):66-80.
[11] ALLEN C, LUO H L, MURDOCK J, et al. Topic modeling the hàn diǎn ancient classics[J/OL]. [2019-06-20]. https://arxiv.xilesou.top/ftp/arxiv/papers/1702/1702.00860.pdf.
[12] NICHOLS R, SLINGERLAND E, NIELBO K, et al. Modeling the contested relationship between Analects, Mencius, and Xunzi: preliminary evidence from a machine-learning approach[J]. The journal of Asian studies, 2018, 77(1): 19-57.
[13] 姜明波. 近十年国内《左传》研究综述[J]. 华夏文化,2013(2):58-61.
[14] 邓勇. 王霸:正义与秩序[D]. 武汉:武汉大学,2007.
[15] 刘巍.《左传》叙战语篇研究[D]. 长春:吉林大学,2013.
[16] 张君蕊.《左传》礼制与"三礼"比较研究[J]. 中国典籍与文化,2017(3):94-109.
[17] 王竹波. 论《左传》"以礼解经"[J]. 现代哲学,2012(4):105-111.
[18] 李佳艺. 从《左传》中探究鲁隐公人物形象[J]. 名作欣赏,2017(17):16-17.
[19] 刘妍彤.《左传》郑庄公人物形象之解析[J]. 文化学刊,2018(3):221-222.
[20] 吕丽,张倩倩,张敬. 浅议《左传》外交辞令的特色[J]. 名作欣赏,2013(17):108-109.
[21] 王立. 婉约有致、辞强不激的语体风格——《左传》外交辞令之探究[J]. 汉字文化,2011(3):55-58.
[22] 许超,陈小荷.《左传》中的春秋社会网络分析[J]. 南京师范大学文学院学报,2014(1):179-184.
[23] BLEI D M, NG A Y, JORDAN M I. Latent dirichlet allocation[J]. Journal of machine learning research, 2003, 3(1): 993-1022.
[24] SCHMIDT B M. Words alone: dismantling topic models in the humanities[J]. Journal of digital humanities, 2012, 2(1): 49-65.
[25] UNDERWOOD T. What kinds of "topics" does topic modeling actually produce[EB/OL]. [2019-06-05]. http://tedunderwood.com/2012/04/01/what-kinds-oftopics-does-topic-modeling-actually-produce/.
[26] GRIFFITHS T L, STEYVERS M. Finding scientific topics[J]. Proceedings of the National Academy of Sciences, 2004, 101(S1): 5228-5235.
[27] 崔凯. 基于 LDA 的主题演化研究与实现[D]. 长沙:国防科学技术大学,2010.
[28] 曲靖野,陈震,胡铁楠. 共词分析与 LDA 模型分析在文本主题挖掘中的比较研究[J]. 情报科学,2018,36(2):18-23.
[29] 杨伯峻,徐提. 春秋左传词典[M]. 北京:中华书局,1985.

作者贡献说明:
何琳:论文选题与框架设计,论文修改;
乔粤:论文撰写与修改,算法实现;
刘雪琪:数据分析与论文修改。

Topic Mining and Evolution Analysis of Social Development in Spring and Autumn Period
——A Case of Studying *Zuo Zhuan*

He Lin Qiao Yue Liu Xueqi
College of Information Science & Technology, Nanjing Agricultural University, Nanjing 210095

Abstract: [**Purpose/significance**] In the context of the rapid development of humanistic computing, this paper uses text mining technology to cluster *Zuo Zhuan*, which provides a reference for quantitative analysis such as topic mining in Spring and Autumn Period, and has a certain reference significance for multi-dimensional reorganization and analysis of classical texts. [**Method/process**] This paper uses text clustering method to analyze *Zuo Zhuan* quantitatively in many dimensions, breaking the linear and chronological record order of *Zuo Zhuan*. Firstly, using the word matching algorithm, the corpus of each vassal state is obtained from the characteristic words of *Zuo Zhuan*. Then the LDA topic model is used to process the characteristic words of *Zuo Zhuan* and the corpuses of selected vassal states. Finally, the topic strength calculation is performed in combination with the time information. [**Result/conclusion**] The experimental results show that the development of the Spring and Autumn Society and the vassal states can be explored according to the theme-word distribution. The development trend of the Spring and Autumn Society and various vassal states can be summarized through the theme intensity curve. Through the LDA topic clustering method, the development of war, politics and diplomacy in the whole society and different vassal states in the Spring and Autumn Period is finally revealed.

Keywords: *Zuo Zhuan* topic mining LDA topic model topic evolution social changes in the Spring and Autumn Period